How to consider fractional counting and field normalization in the statistical modeling of bibliometric data: A multilevel Poisson regression approach

Mutz, Rüdiger ; Daniel, Hans-Dieter

Abstract: The numerical-algorithmic procedures of fractional counting and field normalization are often mentioned as indispensable requirements for bibliometric analyses. Against the background of the increasing importance of statistics in bibliometrics, a multilevel Poisson regression model (level 1: publication, level 2: author) shows possible ways to consider fractional counting and field normalization in a statistical model (fractional counting I). However, due to the assumption of duplicate publications in the data set, the approach is not quite optimal. Therefore, a more advanced approach, a multilevel multiple membership model, is proposed that no longer provides for duplicates (fractional counting II). It is assumed that the citation impact can essentially be attributed to time-stable dispositions of researchers as authors who contribute with different fractions to the success of a publication's citation. The two approaches are applied to bibliometric data for 254 scientists working in social science methodology. A major advantage of fractional counting II is that the results no longer depend on the type of fractional counting (e.g., equal weighting). Differences between authors in rankings are reproduced more clearly than on the basis of percentiles. In addition, the strong importance of field normalization is demonstrated; 60% of the citation variance is explained by field normalization.

DOI: https://doi.org/10.1016/j.joi.2019.03.007

Journal of Informetrics 13 (2019) 643–657

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

How to consider fractional counting and field normalization in the statistical modeling of bibliometric data: A multilevel Poisson regression approach

Rüdiger Mutz[a,*], Hans-Dieter Daniel[a,b]

[a] *Professorship for Social Psychology and Research on Higher Education, ETH Zurich, Zurich, Switzerland*
[b] *Department of Psychology, University of Zurich, Switzerland*

ARTICLE INFO

Article history:
Received 23 July 2018
Received in revised form 12 March 2019
Accepted 12 March 2019
Available online 3 April 2019

Keywords:
Fractional counting
Field normalization
Multilevel model
Poisson regression
Multilevel multiple membership model

ABSTRACT

The numerical-algorithmic procedures of fractional counting and field normalization are often mentioned as indispensable requirements for bibliometric analyses. Against the background of the increasing importance of statistics in bibliometrics, a multilevel Poisson regression model (level 1: publication, level 2: author) shows possible ways to consider fractional counting and field normalization in a statistical model (fractional counting I). However, due to the assumption of duplicate publications in the data set, the approach is not quite optimal. Therefore, a more advanced approach, a multilevel multiple membership model, is proposed that no longer provides for duplicates (fractional counting II). It is assumed that the citation impact can essentially be attributed to time-stable dispositions of researchers as authors who contribute with different fractions to the success of a publication's citation. The two approaches are applied to bibliometric data for 254 scientists working in social science methodology. A major advantage of fractional counting II is that the results no longer depend on the type of fractional counting (e.g., equal weighting). Differences between authors in rankings are reproduced more clearly than on the basis of percentiles. In addition, the strong importance of field normalization is demonstrated; 60% of the citation variance is explained by field normalization.

1. Introduction

Even though there are no binding standards for bibliometric analyses, there are always two central requirements for bibliometric studies in the literature: First, a comparative bibliometric analysis must take into account the different citation cultures of the different scientific disciplines by normalizing bibliometric indicators, in particular citation data (Leydesdorff, Bornmann, Mutz, & Opthof, 2011; Thelwall, 2017; Waltman & van Eck, 2013b). Citation-based bibliometric indicators can be normalized based on citations using a field classification system (source normalization). For example, citation-based raw indicators can be compared with mean values of this indicator in a field (Waltman & van Eck, 2013a), or they can be compared to a field reference distribution in terms of percentiles (Bornmann & Mutz, 2011, 2013). A further normalization strategy

* Corresponding author at: Professorship for Social Psychology and Research on Higher Education, ETH Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland.
E-mail address: mutz@gess.ethz.ch (R. Mutz).

https://doi.org/10.1016/j.joi.2019.03.007
1751-1577/

"... correct[s] for differences in citation practices between fields based on the referencing behavior of citing publications or citing journals" (Waltman & van Eck, 2013b, p. 834) (see also Bornmann and Marx (2015).

Second, a bibliometric analysis is expected to answer the question of how citations are attributed to the authors of papers with multiple authors (Assimakis & Adam, 2010; Bouyssou & Marchant, 2016; Egghe & Rousseau, 1990; Erlen, Siminoff, Sereika, & Sutton, 1997; Kosmulski, 2012; Lindsey, 1992; Perianes-Rodriguez & Ruiz-Castillo, 2015; Perianes-Rodriguez, Waltman, & van Eck, 2016; Radicchi & Castellano, 2011; Tscharntke, Hochberg, Rand, Resh, & Krauss, 2007; van Hooydonk, 1997, p. 944). With the method of *first author counting*, "...only the first author of the N authors is given credit for the multi-authored article..."; ... with *normal or full counting*, "each of the N authors is given full credit, which inflates the number of publications," and with *fractional counting*, "1/Nth of a multiauthored publication is attributed to each of the N authors" (van Hooydonk, 1997, p. 944). Last but not least, with *proportional ranking*, "the author's *rank* on a multi-authored article is an indication of his relative credit" (van Hooydonk, 1997, p. 944). In addition to proportional ranking further methods of weighting can be distinguished, such as harmonic counting (e.g., Hagen, 2010, 2013), geometric counting (e.g., Egghe, Rousseau, & Van Hooydonk, 2000), a counting method that is based on the golden number (Assimakis & Adam, 2010), or last-author counting (Kosmulski, 2012). An overview of counting methods can be found in Waltman (2016). Besides the author-level counting procedures other variants of fractional counting can be distinguished, such as address-level, organization-level, and country-level fractional counting (Waltman & van Eck, 2015, p. 873). Fractional counting is favored for the following reason: "The problem of full counting basically is that co-authored publications are counted multiple times, once for each co-author, which creates a bias in favor for fields, in which there is a lot of co-authorship and in which co-authorship correlates with additional citations" (Waltman & van Eck, 2015, p. 872). Fractional counting avoids this classical problem, but it is also promising regarding bibliometric networks. It "has the attractive property that each action, such as co-authoring or citing a publication, has equal weight" (Perianes-Rodriguez et al., 2016, p. 1192). But it must be noted that fractional counting can lead to distortions that can result in a false picture of the true scientific impact of a single publication. Distortions of that kind were demonstrated in the Leiden Raking (http://www.leidenranking.com/) as an example (Mutz & Daniel, 2015). Gauffriau (2017) has recently published a review of arguments regarding counting methods.

Both numerical-algorithmic methods, fractional counting and field normalization, refer ideally to single publications. In a first step, citations are normalized and/or fractionated. In a second step, the bibliometric data corrected or transformed in this way are then summarized for individual researchers, journals, institutions, or countries, depending on the research question. Statistical aspects, such as random noise, are generally not considered in the procedures. The transformed data, not the raw data, is often the subject of further statistical analyses. The problem then arises as to how these data are to be statistically modeled. Whereas raw citations can be seen as count variables and modeled by means of statistical methods for count variables, crown indicators that normalize citations for field differences are ratios of observed citation counts to expected citations in a field (Waltman & van Eck, 2013a), i.e., real numbers, which are not necessarily normally distributed and are therefore a challenge for the statistical analysis.

What is missing so far are simple statistical approaches for bibliometric raw data that include normalization and fractional counting. This is important, considering that the application of statistical methods is becoming increasingly important in bibliometrics. The aim of this paper is to develop statistical approaches based on count regression models (Ajiferuke & Famoye, 2015; Bornmann & Daniel, 2007, 2016; Mutz, Wolbring, & Daniel, 2017) that take both fractional counting and normalization into account in one statistical model (one-step procedure) primarily for author-level analysis. This also involves a change of perspective. Instead of first correcting and then aggregating the bibliometric data into an indicator (bottom-up) as before, the idea of an indicator should be taken as a starting point (e.g., field normalized citation impact of an institution), and then implications for its correction should be derived (top-down).

2. Statistical modeling of fractional counting and field normalization

2.1. Basic model

Ideally, in bibliometrics, one starts from researchers as authors who publish their research results in a series of publications in scientific journals in co-authorship with other authors. These publications are honored by becoming the subject of other publications (cited reference), which themselves generate citations of the original publication that they cited. It is assumed that citations are not pure chance products but are based on time-stable dispositions (e.g., traits, abilities, skills, competencies) of the researchers that allow them to publish important results and methods that arouse interest in the scientific community in the future as well. Researchers differ in their dispositions to produce excellent research. Ultimately, all personal bibliometric indicators such as the h-index, are based on this assumption. For instance, a university might expect high impact research in future by hiring a researcher with a high h-index.

Bibliometric data has a multilevel structure, as an example with fictitious data with four authors and seven publications show (Table 1): Publications (level-1) are hierarchically assigned to authors and co-authors (level-2). Further, it can be assumed that researchers differ in their overall citation level, and citations of publications by one author are distributed more homogeneously than citations of the combined set of publications by multiple authors. Researchers as authors with high citation levels on average write disproportionally more highly-cited paper than researchers with moderate citation level on the average. If these assumptions were not valid, any comparisons between authors would be pointless. Observed differences between researchers would be purely accidental. Furthermore, the more homogeneous the publication data of

Table 1
Fictitious data WITH duplicates for 4 focused authors and 7 publications, sorted by the publication identifier (PID).

Publication					Fractional counting I			Data		
PID	AUID	AUTHOR	YEAR	DUP	Equal	First	Last	Cit	Fract. Equal Cit	Ref
1	1	Bornmann & Mutz	2010	1	0.50	0.67	0.33	44	22	10
1	2	Bornmann & Mutz	2010	1	0.50	0.33	0.67	44	22	10
2	1	Bornmann	2015	0	1.00	1.00	1.00	29	29	11
3	2	Mutz, Daniel & Cronin	2012	1	0.33	0.50	0.17	375	125	5
3	3	Mutz, Daniel & Cronin	2012	1	0.33	0.33	0.33	375	125	5
3	4	Mutz, Daniel & Cronin	2012	1	0.34	0.17	0.50	375	125	5
4	3	Daniel	2000	0	1.00	1.00	1.00	742	742	5
5	4	Cronin	2010	0	1.00	1.00	1.00	8875	8875	22
6	2	Mutz	2017	0	1.00	1.00	1.00	7	7	1
7	2	Mutz, Singh, Abramo	2011	1	0.33	0.50	0.17	29	9.67	15
7	5	Mutz, Singh, Abramo	2011	1	0.67	0.50	0.83	29	19.33	15

Note. PID = publication identifier, AUID = author identifier (the corresponding author name is underlined in AUTHOR), AUTHOR = authors, YEAR = publication year, DUP = duplicate (0 = no, 1 = yes), Equal = equal weighting, First = first author weighting, Last = last author weighting, Cit = total citations, Fract. Equal Cit = fractionalized citations (equal weighting), Ref = reference value.

authors are, the more that citations of individual publications are dependent on each other (intraclass correlation), and the real sample of independent measurements of researchers decreases (Hox, Moerbeck, & van de School, 2018, p. 5)—a fact that must be taken into account in statistical procedures so as not to generate erroneous statistical inference, since the sample size is included in the calculation of standard errors.

With regard to the scale level, citations are counts. Count data are positive integer values including zero, and they can be assumed to be Poisson distributed (Hilbe, 2014, p. 2). It is characteristic of a Poisson distribution that the expected value or mean value of variable corresponds to its variance. With higher average citation levels of a researcher, it is to be expected that the variability of the citations of a researcher's publications will also increase (the volume of very high and very low cited works). Further, we initially assume a small world in which the authors about whom a statement should be made may also be co-authors of other works that are in the sample of the examined publications. In the case of several authors and co-authors, duplicates of the same publication occur, according to the number of co-authors.

This description of the model can be formalized statistically with a generalized Poisson mixed model as follows (Joe & Zhu, 2005): Given $i = 1$ to N authors with $j = 1$ to J_i publications of each author i, the corresponding total citations for a publication j (level-1) by researcher or author i (level-2), y_{ij}, is Poisson distributed with expected value λ_{ij} (Austin, Stryhn, Leckie, & Merlo, 2018, p. 574):

$$\begin{aligned} & y_{ij} \sim POISSON\left(\lambda_{ij}\right) \\ & \log(\lambda_{ij}) = \beta_0 + u_{0i} \qquad u_{0i} \sim \mathrm{N}(0, \sigma^2_{u0}), \end{aligned} \tag{1}$$

where β_0 denotes the intercept, and u_{0i} denotes the normally distributed random intercept with an expected value of zero and random effects variance of $\sigma^2{}_{u0}$. In this model λ_{ij} is constant for all publications j of a researcher, i.e., $\lambda_{ij} = \lambda_i$. The logarithmically scaled random effect u_{0i} represents the strength of the overall time-stable disposition of a researcher to publish papers with high citation impact. The higher u_i, the higher the citation level of the publications of researcher i is on average. In the case of systematic differences among researchers in the citation level ($\sigma^2{}_u$>0), however, the variance of the observed citations, y_{ji}, no longer corresponds to the expected value or mean value $\exp(\beta_0)$, as it is assumed by the Poisson distribution. Under this condition overdispersed data occurs, where the variance is considerably higher than the mean value. Usually, a negative binomial distribution (NB) deals with such a problem. The NB assumes that u_{0i} is gamma distributed. According to Joe and Zhu (2005, p. 220), the generalized Poisson mixed model (GP) is related to a negative binomial distribution (NB) in the sense that "... the GP distribution can be considered as an alternative Poisson mixture model to the NB distribution for overdispersed count data" (Joe & Zhu, 2005, p. 220). As often occurs in bibliometrics, the overdispersed data are caused not only by extremely highly cited publications but also by the systematic variability between units (e.g., authors, journals, organizations), reflected by the random intercept variance $\sigma^2{}_{u0}$.

The basic count model can be modified to capture rates. Rates relate absolute counts to an exposure — for example, 10 aviation accidents out of 100.000 flights. Rates are important in bibliometrics. Absolute citation counts, for instance, are of limit use due to great differences between fields in the overall citation level. Field normalization relates citations to reference values, which are the expected citations for a field and publication year (e.g., Leiden Ranking, http://www.leidenranking.com/information/indicators). Ultimately, the classical crown indicator is nothing but a rate (Waltman & van Eck, 2013a).

If the rate is defined as count/exposure or citation/reference value, y_{ij}/R_j, then replacing y_{ij} by the rate in Eq. (1), multiplying both side by exposure, R_j, moves it to the right side of the equation. The logarithmically transformed reference value (or exposure) is then included in the regression model as an offset (Burrell, 2007, p. 17), as follows (Austin et al., 2018, p. 574):

$$y_{ij} \sim POISSON\left(\lambda_{ij}\right)$$
$$\log(\lambda_{ij}) = \beta_0 + u_{0i} + \log_e\left(R_j\right) \quad u_{0i} \sim \mathrm{N}(0, \sigma^2_{u0}), \tag{2}$$

where $\exp(\beta_0+u_{0i})$ is the overall field-independent citation rate for researcher *i*, and $\log(R_j)$ is the logarithmically transformed reference value, R_j, for publication *j* (Raudenbush & Bryk, 2002, p. 309). The offset is nothing but a predictor variable with a fixed regression coefficient of 1.0. For example, if a researcher *i* has an overall citation rate of $\exp(\beta_0+u_{0i}) = 10$ and reference value of $R_j = 10$ for a single publication *j*, then his or her citation rate in units of publication *j* amounts to 100 citations $(=\exp(\beta_0+u_{0i}+\log_e(R_j)) = \exp(\log(10)+\log(10)) = 100)$. By multiplying the citation rate with the reference value, one obtains a citation rate in units of the respective field of the selected publication.

The citation window in time units can provide an offset as well. The intercept-only model (Eq. (2)) can be extended by a regression model with a set of *k* predictors on the level of individual publications ($\Sigma\beta_k x_{kij}$) and/or on the level of individual authors ($\Sigma\beta_k x_{ki}$). The expression $\exp(\beta_0+u_{0i})$ represents the expected field normalized trait value of an author *i* scored in citation units.

The classical fractional counting method can also be implemented using an offset. In fractional counting, a portion of the citations (e.g., equal weighting) is allocated to the respective author and co-authors of a publication, whereby the proportions of all authors of a publication add up to 1.00. In fractional counting, the citation score is therefore multiplied by the fraction. In terms of the count regression model, the offset is defined as the logarithmically transformed inverse of the fraction of the author, $\log_e(1/p_{ij})$, where p_{ij} is the fraction of author *i* of publication *j*. In both fractional counting and field normalization, the logarithmically transformed product of the fraction and the reference value ($\log_e(R_j\ 1/p_{ij})$) serves as an offset.

The example using fictitious data (Table 1) shows for four scientists and seven publications the data for an analysis that can consider fractional counting and field normalization. Three types of fractional counting are distinguished, which are used in a somewhat different sense than in the literature (van Hooydonk, 1997): equal weighting or "fractional counting" (EQUAL), first author weighting (FIRST), and last author weighting, both variants of "proportional ranking" (LAST). Whereas for equal weighting, the weight for each author is determined as 1/number of authors, for first author weighting, the actual author rank is divided by the rank sum. When this fraction is subtracted from 1.0, the weight is obtained ($1-(\text{rank}_{\text{Author}}/\text{rank sum})$). For instance, the rank sum for the first publication (PID = 1) is 1 + 2 = 3 (Table 1). The first author (rank 1) thus receives the weight (1 - (1/3)) = 0.67. In the last author ranking, only the ranking is rotated. The last author gets rank 1, the first author gets rank n (n = number of authors). If field normalization and fractional counting are considered, the offset (Eq. (2)) for the first publication of the first author ("Bornmann") would be equal ($\log_e(R_j\ 1/p_{ij}) = \log_e(10 \times 1/0.67) = 2.70$). There are two reasons for rhe selection of this set of counting procedures: First, "proportional ranking" (LAST, FIRST) considers the author's rank on a multi-authored paper, which indicates the relative merit of an author‘s contribution. Tscharntke et al. (2007) identified different cultures of ranking authors in multi-authored papers, such as the "sequence-determines-credit" approach, or the "first-last-author emphasis". Boyer, Ikeda, Lefort, Malumbres-Olarte, and Schmidt (2017) proposed a counting procedure based on the contribution percentage provided by the authors themselves. "Equal weighting" is included, because it is one of the most common fractional counting procedure. Second, for statistical reasons it is important to compare quite different procedures with fractions greater than 0 in order to check, whether the type of fractional counting really matters.

2.2. Multiple membership model

Even if the approach presented can solve the problem of field normalization and fractional counting, duplicates of publications remain in the data set. It would be desirable to have an approach that models the citations for each publication without having to resort to duplicates. Also, an approach in which the type of fractional counting (equal, first author, last author) does not matter would be of interest. These problems can be solved by a *multiple membership* approach within the framework of multilevel modeling (Cafri, Hedeker, & Aarons, 2015; Chung & Beretvas, 2012; Fielding & Goldstein, 2006; Goldstein, 2011a, pp. 255–256; 2011b; Gotthard & Calzolari, 2017; Tranmer, Steel, & Browne, 2014).

It is assumed that first authors and co-authors, with their stable dispositions (e.g., skills, abilities, traits) have a significant influence on the resonance of a publication in the scientific community, which is expressed in the number of citations that a publication receives. However, a single publication and its citation cannot make a reliable statement about a person. This requires replications via several publications of the respective author, which are often written in cooperation with other authors. A publication and the citations associated with it are thus assigned simultaneously to the respective author and to several co-authors in the case of multiple-author papers. If authors are regarded numerically as categories, the publication belongs to several categories (multiple membership) with different weighting (Fig. 1). Weights determine the extent to which each author has contributed to the paper. These weights are nothing more than the fractions within the different fractional counting procedures (e.g., equal weighting), which add up to 1.0 for a publication (Cafri et al., 2015, p. 411).

With regard to the example (Table 1), the data are organized in such a way that the information for the individual publications is depicted in the rows; the four authors with their respective fractions, i.e., proportions of the respective publication, are depicted in the columns next to the dependent variable (citations), (Table 2).

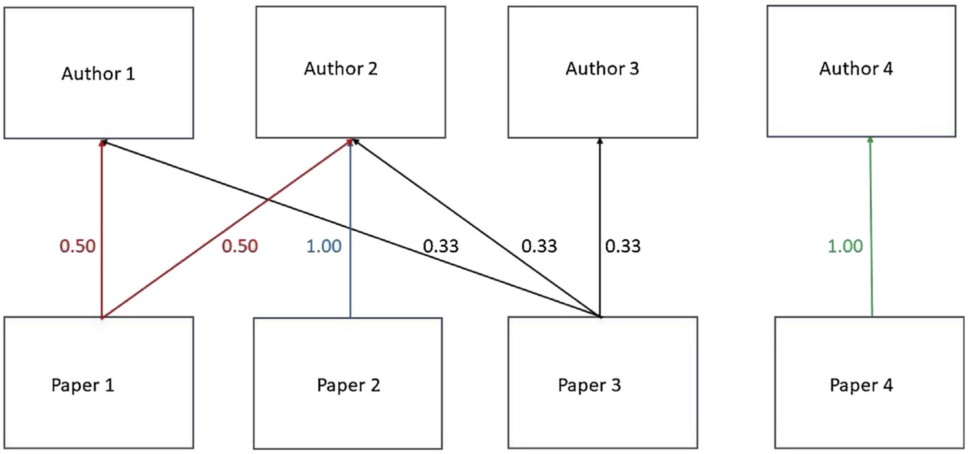

Fig. 1. Multiple membership design.

Table 2
Fictitious data WITHOUT duplicates for 4 focused authors and 7 publications of Table 1.

	Publication	Fractional counting II (equal weighting)					Data	
		AUID						
PID	YEAR	1 (Bornmann)	2 (Mutz)	3 (Daniel)	4 (Cronin)	5 (Other)	Cit	Ref
1	2010	0.50	0.50	0.00	0.00	0.00	44	10
2	2015	1.00	0.00	0.00	0.00	0.00	29	11
3	2012	0.33	0.33	0.00	0.33	0.00	375	5
4	2000	0.00	0.00	1.00	0.00	0.00	742	5
5	2010	0.00	0.00	0.00	1.00	0.00	8875	22
6	2017	0.00	1.00	0.00	0.00	0.00	7	1
7	2011	0.00	0.33	0.00	0.00	0.67	29	15

Note. PID = publication identifier, AUID = author identifier, YEAR = publication year, Cit = total citations, Ref = reference value.

The multilevel multiple membership model is different from the basic model for fractional counting I (Eq. (2)). The random intercept u_{0i} of an author *i* is weighted with the respective fraction p_{ij} for a single publication *j* (Fig. 1) and summed over all authors of a publication as follows:

$$y_j \sim POISSON(\lambda_j)$$
$$\log(\lambda_j) = \beta_0 + \sum_{i=1}^{N} p_{ij} u_{0i} + \log(R_j) \quad u_{0i} \sim N(0, \sigma_{u0}^2). \tag{3}$$
$$\text{with} \sum_{i=1}^{N} p_{ij} = 1.00$$

Using this multiple membership approach, the disposition score (i.e., random effect u_{0i}) for each author can be determined without assuming any duplicates in the data. The network of the first author and the co-authors is represented by the weights (Tranmer et al., 2014; Tranmer, Pallotti, & Lomi, 2016). In the basic model of fractional counting (see Section 2.1 above) the logarithmic value of p_{ij}, $\ln(1/p_{ij})$, enters the equation; in the multiple membership model, the non-logarithmic version is used (Eq. (3)) as the weight of the person parameter u_{0i}. In contrast to the numerical method of fractional counting, in which the data is first transformed and then aggregated to the level of persons or organizations (bottom up), the multi-membership model is based on dispositions of the authors that affect the citations depending on the contribution of the author to the publication—that is, his or her fraction (top-down).

Regarding the data example, for the four authors (Bornmann, Mutz, Daniel, Cronin××××) the following random intercept values u_{0i} are estimated: −2, −1, 2, and 3 and the intercept b_0 of 3. Conversely, this information can be used to predict the observed citation of a publication according to Eq. (3). The citation value for the first publication with authors Bornmann ($p_{11} = 0.5$) and Mutz ($p_{21} = 0.5$) with reference value $R_1 = 10$ equals $y_1 = \exp(b_0 + \Sigma p_{j1}\, u_i + \log(R_1)) = \exp(3 + 0.5 \times -2 + 0.5 \times -1 + \log_e(10)) = 44.8$. The real citation value of 44 (Table 2) is slightly overestimated by 0.8 citations.

2.3. Addendum: authors not in the sample, intraclass correlation, and ranking

Three questions remain unanswered: First, how are first authors and co-authors to be treated about whom no statement is to be made in a bibliometric analysis, i.e., who are not part of the sample of researchers? Second, how can it be checked that the observed variability between the researchers with respect to their mean citation impact actually reflects systematic variance and not just random noise? Third, how can researchers as authors be compared with each other?

There are several answers to the *first question* about first authors and co-authors who are not part of the sample of researchers in the study. First, only the members of the sample, provided they are authors, could be taken into account in the weighting. One variation would be to determine the weighting of all authors of a contribution, simply omit the group of authors and co-authors not in the sample in their weight, so that overall the sum of the weights becomes less than 1. However, the contribution of the members of the sample is thus rather overestimated: “This restriction on the weights isn't a necessary modeling constraint, but ... is consistent with respect to how multiple membership models are often parameterized” (Cafri et al., 2015, p. 412). Second, an additional line (Table 1, PID = 7) or column (Table 2) labeled “Other” with Author ID (AUID) could be added for each publication, taking into account the total of first authors and co-authors who are not the subject of the study and their total weight. Thus, the author “Mutz” receives the weight 0.67 (“First author weighting”) for publication 7 and the other two authors a total value 0.33. There are as many “Other” groups as there are publications that contain first authors or co-authors who are not the subject of the study.

The *second question* aims at intraclass correlations (Mutz, Bornmann, & Daniel, 2012, p. 3). An intraclass correlation ρ indicates the proportion of systematic variance of mean citation levels between authors to the total variance of citations across all publications and varies between 0 (= random noise) and 1 (perfect discrimination). The higher the ICC, the more homogeneous the publications of an author in the sample are with regard to their citations. In this case, the data are no longer independent. Therefore, the sample of independent units decreases with strong impact on the standard errors of the estimated parameters, which increase (more inaccuracy). For a Poisson regression, the ICC can be calculated using the following formula (Austin et al., 2018, p. 575; Raabe-Hesketh & Skrondal, 2012, p. 697):

$$\rho = \frac{\exp(2\boldsymbol{\beta}\mathbf{X} + 2\sigma^2) - \exp(2\boldsymbol{\beta}\mathbf{X} + \sigma^2)}{\exp(2\boldsymbol{\beta}\mathbf{X} + 2\sigma^2) - \exp(2\boldsymbol{\beta}\mathbf{X} + \sigma^2) + \exp(\boldsymbol{\beta}\mathbf{X} + \sigma^2/2)}, \tag{4}$$

where the expression $\exp(\boldsymbol{\beta}\mathbf{X}+\sigma^2/2)$ reflects the within-variance of citations of publications of an author and σ^2 the random intercept variance σ^2_{u0} (Eqs. (1)–(3)). The term $\boldsymbol{\beta}\mathbf{X}$ represents the Poisson fixed-effects part of the mixed-effects model. Therefore, except for a model without predictors ($\boldsymbol{\beta}\mathbf{X} = \beta_0$), the intraclass correlation varies with different values of the predictors ($\mathbf{X}$). For the multiple membership model (fractional counting version II), instead of σ^2_{u0} the variance of the weighted sum of random effects ($\Sigma p_{j1}\, u_i$) for each publication is used.

Another measure is the median ratio rate (MRR), which measures the median relative change in the occurrence of citations when comparing publications from two randomly selected different authors that are ordered by citation rate (Austin et al., 2018, p. 576; Raabe-Hesketh & Skrondal, 2012, p. 697).

$$MRR = \exp(\sqrt{2\sigma^2}\phi^{-1}(0.75)), \tag{5}$$

where ϕ^{-1} is the inverse of the standard normal distribution function and σ^2 the random intercept variance σ^2_{u0}.as in Eq. (4).

For example, at an MRR of 5.0, a random publication of an author A would have 5 more citations in the median compared to a random publication of an author B, who has a lower rank in the citation level than author A.

Regarding the *third question* of author comparisons, the estimates of the random intercept, u_i (in Eqs. (1)–(3)), can be used, which are called empirical Bayes estimates (Hox et al., 2018, p. 244): Basically, empirical Bayes (EB) estimates derive from two concepts of the expected value of a researcher‘s mean value: If there is no reliable information about a researcher *i*, the overall intercept β_0 (Eq. (2)) or mean value across all researchers is the best estimate for the mean value of a researcher *i*. Vice versa, if there is very reliable information about a researcher, then the individual intercept value of an Ordinary Least Squares (OLS) regression for researcher *i*, $\beta^{OLS}_{0i,}$ is the best estimate. The EB estimate β^{EB}_i for researcher *i* combines both concepts by weighting them with the reliability, as follows (Hox et al., 2018, p. 244):

$$\beta^{EB}{}_i = \lambda_i\beta^{OLS}{}_{0i} + (1-\lambda_i)\beta_0, \tag{6}$$

where λ_i is the reliability of the OLS regression intercept β^{OLS}_{0i} varying between 0 (=not reliable) to 1 (=perfectly reliable). The reliability is a function of the random intercept variance σ^2_{u0}, the within variance (Eq. (4)), and the sample size n_i (i.e., number of publications of a researcher *i*). The higher the random intercept variance, the lower the within variance, and the higher the sample size, the higher the reliability of the OLS intercept β^{OLS}_{0i}. The higher the reliability, the more the EB estimates shift to the estimated OLS intercepts. “As a result, the regression coefficient are *shrunk* back towards the mean coefficient for the whole data” (Hox et al., 2018, p. 244). Although the EB are biased, they are closer to the true values than any other kind of estimates.

Table 3
Sample description.

Variable	*N*	*M*	*SD*	*VAR*	*MIN*	*25%*	*Mdn*	*75%*	*Max*
Publication level									
Total citations	5,095	14.62	85.79	7360.1	0	1	4	13	5491
NCS	4,494	1.14	3.71	13.77	0	0.14	0.54	1.28	207.06
Ptop10p	5,095	0.13	0.33	0.11					
Author level									
Gender (1=female, 0=male)	254	0.28	0.45	0.20					
Years since the first publication	254	14.78	9.88	97.71	1	7	12	20	37
Total citations	254	381.53	774.3	$6.0\ 10^6$	0	9	70	285	7,932
Fractionated citation[a]	254	4.79	7.08	50.2	0	1.12	2.59	5.99	69.72
Field normalized fractionated citation[a]	254	0.43	0.41	0.17	0	0.17	0.32	0.54	2.91
MNCS	254	23.59	40.95	1,677	0	2.00	8.94	28.53	332.88
PPtop10p	254	2.81	5.08	28.78	0	0	1	4	34

Note. NCS = normalized citation score, Ptop10p = paper in top 10% percentile or not, PPtop10p = number of articles in the 10% percentile; MNCS = mean normalized citation score.

[a] The fraction counting was done with equal weights for each author.

3. Data and methods

3.1. Data

For the sample, we used the members of a quantitative methodology section of an undisclosed prominent German academic society for social sciences. The members were listed in the society's membership directory of 2016 (N = 289). An information center for the German-speaking countries made a database available with information on the publications of authors and the scientific careers of the selected sample. This database and also Google were used to collect additional information on the members of the section (e.g., affiliation, gender).

As the bibliographic database, Scopus was used. The Scopus Author Identifier and publication list of each member were retrieved using the author search. As the lists are created automatically by Scopus and can therefore contain errors, where possible we additionally checked the lists against the members' websites. Members who had no publications (n = 35) were eliminated from the sample, so that the final sample available for the development of the scale was N = 254 persons. Overall, N = 5,095 publications were considered in the data analysis. 29.1% of the publications were written by a single author, 27.2% by two authors, 18.9% by three authors, and 24.9% by more than three authors. First authors and co-authors who were not part of the sample were summarized as "Other" for each publication. This resulted in N = 3,678 "other" authors.

The Centre of Science and Technology Studies B.V. (CWTS) at Leiden University provided us with bibliometric indicators by matching the set of articles with bibliographic data in Web of Science according to the identified accession number for each article (UT codes).

3.2. Variables

The following bibliometric data were available for the sample of publications (Table 3): raw total citation counts, normalized citation scores (NCS), publication in the 10% percentile of a field or not (Ptop10p). The fields were defined according to the subject categories of Web of Science, which are created by assigning journals to one or more subjects. On the aggregate level (author) mean normalized citation scores (MNCS) and 10th percentiles (PPtop10p) were calculated.

Additionally, age and gender were included in the analysis (Table 3). Since the researcher's age was usually not available, the researcher's academic age was used, defined as the number of years since the researcher's first publication (= 2016 - year of first publication). Fractional counting was done for the raw citation scores as well as for the field-normalized citations scores. Three kinds of fractions in the fractional counting were generated: equal weighting, first author weighting, and last author weighting.

3.3. Statistical procedures

For the classical fractional counting (fractional counting I), the data were organized comparable to Table 1 with duplicates for first authors and co-authors who were members of the sample (additional duplicate for "other" authors). For multiple membership (fractional counting II), the data were organized as in Table 2 without any duplicates. The statistical analyses (multilevel Poisson regression) were done using the procedure PROC GLIMMIX provided by the statistical software package SAS (SAS Institute Inc., 2011, p. 2805), which allows the estimation of multilevel multi-membership models as well. To compare the models, the Schwarz-Bayes information criterion (BIC) was used. The lower the BIC, the better the model is. Due to the fact that PROC GLIMMIX used the residual pseudo-likelihood technique in the estimation process, only pseudo-BIC were presented, which are of limited use. As another fit measure in Poisson regression analysis, the ratio of the χ^2 value and the degrees of freedoms could be used. In the case of χ^2/df of 1.0, the model fits the data well. For

Table 4
Model comparison.

Model			Raw citation data						Field-normalized citation data						
No	Distribution	Description	BIC	χ^2/df	σ^2_u	σ^2_{uw}	ICC	MRR	BIC	χ^2/df	σ^2_u	σ^2_{uw}	ICC	MRR	R^2
0	POISSON	Intercept only	427,170	537.5					299,907	235.1	–		–	–	–
1	POISSON	Random intercept	518,018	55.6	2.35	–	,99	4.3	285,855	30.2	1.06	–	.70	2.7	.55
2	POISSON	Fraction I-Equal	747,192	80.5	2.51	–	.99	4.5	366,388	38.9	1.18	–	.63	2.8	.53
3	POISSON	Fraction I-First	2520,002	273.8	2.61	–	.99	4.7	1092,355	117.9	1.29	–	.71	3.0	.51
4	POISSON	Fraction I-Last	699,579	75.3	2.61	–	.99	4.7	341,483	36.1	1.25	–	.60	2.9	.52
5	POISSON	Fraction II-Equal	65,422	10.8	5.84	1.15	.91[a]	2.8[a]	49,415	7.9	2.75	0.46	.31[a]	1.9[a]	.61[a]
6	POISSON	Fraction II-First	65,059	10.7	3.87	1.05	.92[a]	2.7[a]	49,373	7.9	1.80	0.92	.31[a]	1.9[a]	.60[a]
7	POISSON	Fraction II-Last	67,000	11.4	10.99	1.46	.92[a]	3.2[a]	50,386	8.0	5.19	0.51	.32[a]	2.0[a]	.65[a]
8	BINARY	Fraction II-Equal	–	–	–	–	–	–	25,230	–	1.07	0.15	.04	–	–
9	BINARY	Fraction II-First	–	–	–	–	–	–	25,039	–	0.91	0.12	.04	–	–
10	BINARY	Fraction II-Last	–	–	–	–	–	–	25,461	–	1.04	0.17	.05	–	–

Note. No = model identifier, BIC = Pseudo Schwarz Bayes Information Criterion (BIC) except for No = 0 (real BIC), χ^2/df = ratio of χ^2 and degrees of freedom (fit statistic), σ^2_u = variance of random effects of authors, σ^2_{uw} = variance of weighted random effects of authors, ICC = intra class correlation, MRR = median rate ratio, R^2 = amount of variance of σ^2_u explained by the field normalization factor (offset). BINARY refers to the percentiles (PPtop10p).

[a] Calculation is based on the weighted random effects (σ^2_{uw}).

PPtop10p, a binary variable (classified as top 10% publication or not), a multilevel logistic regression model was applied (Bornmann, Mutz, & Daniel, 2013; Bornmann, Stefaner, de Moya Anegon, & Mutz, 2014; Bornmann, Stefaner, de Moya Anegon, & Mutz, 2015). The ICC for binary variables is calculated as follows (Hedeker, 2003, p. 1439): $\rho = \sigma^2/(\sigma^2 + \pi^{2/3})$. The program code for the Poisson regression model including the data of the example (Table 2) is provided in the supplementary material.

4. Results

4.1. Model comparison

In all, 11 models were calculated, one for raw citation data and one for field-normalized citation data (Table 4). A model that contains only one intercept (M_0) is distinguishable from a model that contains additionally one random intercept for each author (M_1). Further, models with classical fractional counting I (M_2-M_4) are distinguished from models with multiple-membership fractional counting II (M_5-M_7). Last but not least, fractional counting models can also be estimated for the percentiles (PPtop10p), which already contain a field normalization and therefore cannot be applied to raw data (M_8-M_{10}). Since the focus of this contribution is on the analysis of raw data, the binary model with percentiles was only calculated for fractional counting II as an add-on to the models for raw citation data. Much information and statistical power is lost by dichotomizing a continuous variable (e.g., Altman, 2006).We obtained the following results:

- *Multilevel model*: By including random intercepts (without fractional counting) the count model was significantly improved. Thus, the ratio χ/df for the random intercept model (M_1) compared to the intercept-only model (M_0) decreased irrespective of whether or not the data are field normalized (e.g., from 537.5 to 55.6 for raw citation data). Overall, the model fit in all models with a χ/df ratio was much higher than 1.0 and therefore not optimal. However, it must be taken into account that no explanatory variables other than field normalization or fractional counting were taken into account in the models ("empty model").
- *Field normalization*: When the estimated random intercept variances (σ^2_u, σ^2_{uw}) for raw citation data and field normalized citation data were compared, more than 50% of the variance in fractional counting I and more than 60% of the variance in fractional counting II was explained by field normalization (R^2). Overall, a large part of the observed variability in the citations was thus reduced by differences in the citation level of the fields in which the authors had published. This should be seen against the background that the sample was a relatively homogeneous group of social scientists specialized in quantitative methodology in the social sciences (restricted range of fields). This result underlines the great importance of field normalization in bibliometrics.
- *Intraclass correlation*: The intraclass correlation (ICC) with values over .60/.90 for raw citation data and field normalized citation data in the case of fractional counting I was very high. Whereas the ICC for fractional counting II was even high for raw citation data (~.90), it was lower for field-normalized citation data. Overall, the ICCs reflected systematic variance between authors beyond random fluctuations in all models. Comparisons among authors were therefore justified. However, the ICC also included the "others" group, which were first authors and co-authors not included in the sample.
- *Comparison fractional counting I and II*: Whereas for fractional counting I the values for BIC and the ratios, χ^2/df, varied strongly, the corresponding values for fractional counting II varied only slightly both for raw and field-normalized citation data. This can be taken as an indication that fractional counting I was more dependent on the type of counting procedure (e.g., equal) than fractional counting II. Even if the random intercept variance, σ^2_u, showed greater fluctuations for fractional

Table 5
Correlations (Pearson) among the estimated random effects for multilevel models for raw and field-normalized data, differently fractionalized, and raw citation scores (N = 254 researchers).

		Raw citation data							Field-normalized citation data						
No	Description	1	2	3	4	5	6	7	1	2	3	4	5	6	7
POISSON															
1	Random intercept	1.0	.93	.74	.93	.86	.82	.87	1.0	.86	.58	.86	.83	.80	.83
2	Fraction I-Equal	-	1.0	.90	.98	.80	.76	.82	-	1.0	.85	.94	.72	.70	.73
3	Fraction I-First	-	-	1.0	.82	.64	.59	.65	-	-	1.0	.63	.52	.51	.54
4	Fraction I-Last	-	-	-	1.0	.80	.76	.83	-	-	-	1.0	.70	.67	.72
5	Fraction II-Equal	-	-	-	-	1.0	.98	.98	-	-	-	-	1.0	.98	.98
6	Fraction II-First	-	-	-	-	-	1.0	.93	-	-	-	-	-	1.0	.94
7	Fraction II-Last	-	-	-	-	-	-	1.0	-	-	-	-	-	-	1.0
BINARY															
8	Fraction II-Equal	.37	.28	.24	.26	.34	.35	.31	.61	.52	.35	.49	.56	.57	.54
9	Fraction II-First	.34	.26	.22	.24	.33	.35	.31	.58	.49	.34	.47	.56	.58	.53
10	Fraction II-Last	.38	.30	.25	.27	.34	.34	.31	.63	.53	.35	.50	.55	.55	.53
MANIFEST															
11	Citations	.51	.43	.28	.44	.46	.46	.44	.43	.33	.10	.38	.33	.33	.33
12	Log_e(cit)	.84	.72	.51	.72	.75	.72	.75	.56	.39	.14	.50	.48	.44	.52
13	MNCS	.31	.26	.16	.26	.28	.28	.27	.44	.40	.25	.40	.42	.42	.42
14	PP_top10p	.22	.17	.11	.16	.17	.17	.16	.34	.30	.17	.29	.31	.31	.31
15	Gender	-.14	-.10	.01	-.14	-.09	-.09	-.08	-.04	.04	.15	-.05	.04	.04	.03
16	Acad. Age	.46	.51	.39	.53	.40	.37	.42	-.13	.10	-.23	.03	-.17	-.18	-.13

Note. The grey shaded regions indicate the inter-correlations among different fractional counting methods, separated for raw and field-normalized data.

counting II than for fractional counting I, these differences were somewhat leveled when the weighted random effects variance, σ^2_{uw}, was considered.

- *Comparison to PPtop10p*: It became clear that the random intercept variance, σ^2_u, and the ICC of ~.04 for PPtop10p were significantly smaller than for fractional counting I and II. The dichotomization into excellent publications (i.e., publications in PPtop10p) and non-excellent publications was accompanied by a lower differentiation of the researchers than in models that referred to raw citations scores.

4.2. Correlations among random effects variables with and without fractional counting

According to Eqs. (1)–(3), random effects *u* can be estimated for each of the models (Table 4) on the level of researchers. In addition to the model comparison, direct correlations under the random effects can provide information about similarities and differences between the individual fractional counting and field normalization methods (Table 5, highlighted in grey). A further exploratory factor analysis (Table 6) can reveal underlying dimensions (e.g., fractional counting I and II or field normalization or not).

Overall, the random effects were generally highly correlated among themselves for both fractional counting I and fractional counting II (>.80) (Pearson correlation for logarithmic scaled data, Table 5, row 1–7). Nevertheless, the random effects for fractional counting II were more similar and more strongly correlated among themselves (>.92) than

Table 6
Factor loading matrix of exploratory factor analyses of differently fractionalized data (VARIMAX-rotated).

MNo	Description	Factor 1	Factor 2	Factor 3	h^2
Raw data					
1	Random intercept	**.89**	.30	.21	.92
2	Fraction I-Equal	**.90**	.12	.40	.98
3	Fraction I-First	**.73**	−.01	.59	.89
4	Fraction I-Last	**.92**	.11	.32	.95
5	Fraction II-Equal	**.86**	.47	.01	.96
6	Fraction II-First	**.81**	.52	−.01	.92
7	Fraction II-Last	**.88**	.41	.02	.94
Field normalized data					
1	Random intercept	.30	**.71**	.51	.85
2	Fraction I-Equal	.29	.52	**.78**	.96
3	Fraction I-First	.03	.29	**.89**	.87
4	Fraction I-Last	.43	.49	**.67**	.87
5	Fraction II-Equal	.25	**.92**	.27	.98
6	Fraction II-First	.22	**.92**	.24	.95
7	Fraction II-Last	.30	**.89**	.28	.96
Variance explained by each factor		5.69 (40.6%)	4.33 (30.9%)	2.96 (22.1%)	

Note. Factor loadings greater than .60 are shown in bold face; h^2 = communality.

for fractional counting I, regardless of whether they had been field normalized or not. This indicates that fractional counting II no longer depended on the type of fractional counting, i.e., the procedure was relatively robust against different weighting of the authors' contributions. However, fractional counting I and II were not perfectly correlated and varied to some degree (.80, .59), indicating that the two methods were not redundant. Similarly, the random effects of raw and field-normalized citation scores correlated across all fractional count procedures between .60 and .70 (not reported). This also indicated that field normalization in favor of raw citation data cannot be dispensed with.

An exploratory factor analysis can provide information about the correlations. Using principal component analysis, the factors were selected according to the eigenvalue criterion (variance of factor at least 1) and the scree test. A factor solution is, however, not unique. Any rotation of the loading matrix $\mathbf{T}$ can generate a loading matrix $\mathbf{A}^*(=\mathbf{AT})$, which equally fits the observed correlation matrix $\mathbf{R}$ as the original loading matrix $\mathbf{A}$: $\mathbf{R} = \mathbf{AT(AT)'=A^*A^{*'}=AA'}$. We therefore adopted the criterion of a simple structure in the sense that the loading matrix contained a lot of high and low loadings (0/1). A simple structure was approximated by a rotation (VARIMAX), which maximizes the variance of the loadings for each factor (Table 6). According to the eigenvalue criterion and scree test, three factors were extracted, which explained 5.69 + 4.33 + 2.96 = 12.98 of the total variance of 14 (= number of variables), i.e., 92.7% of the total variance was explained by the three factors, a quite optimal fit. The first factor with 40.6% explained variance represented fractional counting I and II applied on raw citation data. Roughly speaking, the second and third factor represented the field-normalized data with the distinction between fractional counting II (factor 2) and fractional counting I (factor I). The loadings for the fractional counting II random effects were more homogenous (factor 2) than for the fractional counting I (factor 3).

4.3. Correlation of latent variables with other covariates

So far, only the correlations among the various random effects variables were considered. Of interest may also be how these variables correlate with the random effects estimated from the percentiles, the observed citations data, and, last but not least, the sociodemographic variables. Field-normalized observed citations scores were expected to correlate strongly with the corresponding field-normalized random effects (Table 6, rows 8–10, rows 13–14). And the other way around, non-normalized citations with fractional counting were expected to correlate strongly with the corresponding random effects with fractional counting based on non-normalized citations (Table 5, rows 11–12). These expectations were also confirmed in the empirical analysis.

Thus, the random effects estimated from percentile data (binary, rows 8–10) were more highly correlated with the random effects based on field-normalized citation data (between .34 and .63) than with those based on raw citation data (between .22 and .37).

The percentiles (PPtop10p, row 14) that were actually field-normalized correlated more highly with the random effects that were based on field-normalized data than with those based on raw citation data. Nevertheless, the correlations were only low (∼.30).

As expected, the raw citations (row 11) correlated more highly with the random effects without field normalization than with the random effects with field normalization, regardless of the fractional method. Since the random effects were logarithmically scaled, the correlations with the $\log_e(\text{cit})$ were usually higher (row 12).

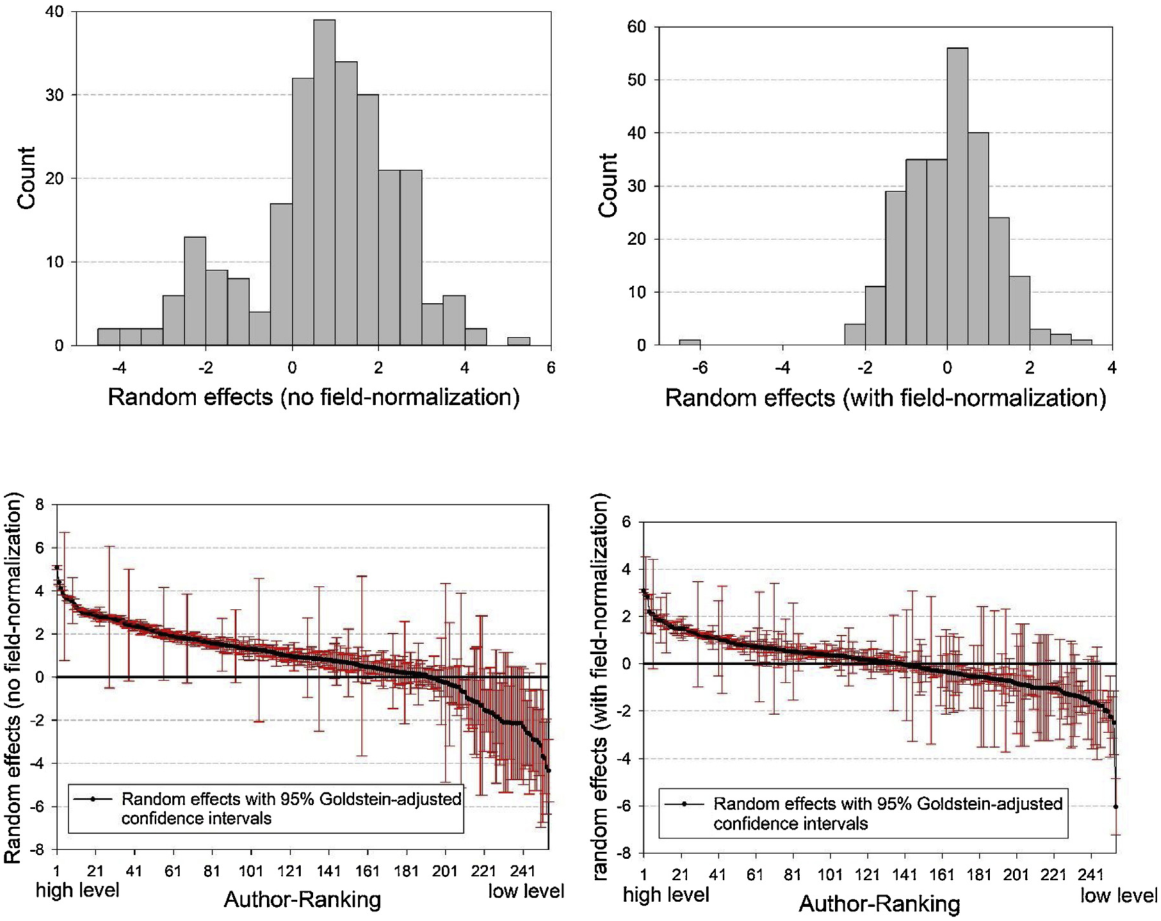

Fig. 2. Histograms of random effects for fractional counting II (above) and scatter plots of random effects for authors (below) from left (high citation level) to right (low citation level) with Goldstein-adjusted confidence intervals without field normalization (left) and with field normalization (right).

With regard to sociodemographic variables (gender, academic age) there was no significant influence of gender over all random effects. A moderate correlation of academic age with random effects was observed only for random effects without field normalization.

4.4. Ranking of authors

Within the framework of evaluative bibliometrics, it is always of interest to rank people, institutions, and countries. For example, for the random effects of each author for fractional counting II (equal weighting) both with and without field normalization, a histogram and a ranking with confidence intervals were generated (Fig. 2).

The histograms show that, contrary to the expectation of highly skewed data in bibliometrics, the values were approximately normally distributed, which more or less results from the logarithmically scaled random effects. The non-field normalized values even showed a bimodal distribution (mixture distribution). With field normalization, the bimodal distribution disappeared, the variance shrank, and an outlier became visible.

The histograms say something only about the distribution, but the scatter plots (Fig. 2, below) provide some information about the accuracy of each estimated random intercept. The accuracy was represented by the 95% confidence interval, which is Goldstein-adjusted (1.396*standard error). Additionally, the more unreliable a random intercept is (e.g., small sample size), the more it is shifted to the overall mean value (empirical Bayes estimates). For non-overlapping confidence intervals, an effect beyond chance can be assumed (Goldstein & Healy, 1995, p. 176). The higher the confidence interval, the more inaccurate the parameter estimation is. The authors were ranked with decreasing intercept value, i.e., the citation level increased from left to right. It became clear that many more authors in the lower range (< 0) had larger confidence intervals than authors in the upper range had. The estimate of random effects of the authors in the lower range of the ranking were therefore much less precise than estimates in the upper range. The strongest difference was between persons who had values around 0 and persons who had values greater than 1.0 (non-overlapping confidence intervals), regardless of whether field normalization was present or not.

5. Discussion

Field normalization and fractional counting are often mentioned as indispensable numerical-algorithm procedures in bibliometrics. Ultimately, these are bottom-up methods in which the raw data are first transformed and then further analyzed. There are three decisive advantages of these procedures. Transformed data can be generated without having to consider the later purpose of the analysis in advance. Data can be provided with field normalization and fractional counting in order to be aggregated to the level of persons, institutions, or journals, depending on the research question. Furthermore, duplicates in the data record are no problem. Due to fractional counting, the number of publications and the number of citations during aggregation remains constantly comparable to a data set containing only one publication at a time. Fields in which co-authorship is very common or in which co-authorship is correlated with citations are not advantaged (Perianes-Rodriguez & Ruiz-Castillo, 2015, p. 975)). Third, the procedures can be applied to all data, whether a publication has 100 authors or only 1. All publications are treated in the same way.

A decisive disadvantage of the common procedures of field normalization and fractional counting is that they do not consider random noise and measurement error (accuracy of measurements). For instance, an author A may have 5 publications, and another author B may have 100 with citations for each publications. The mean citation value for author B with 100 publications might be much more accurate than the mean value for author A.

With fractional counting I (a multilevel Poisson regression), an approach was presented that could take fractional counting and field normalization into account in *one* statistical model using common statistical methods that are included in most statistical software products (SAS, STATA, R). An offset variable with the reference value or fractional counting is simply built into the Poisson model. Nevertheless, the procedure is not optimal, since duplicates in the data set must continue to be used with unknown effects on the statistical model estimation.

For this reason, a further approach, fractional counting II, based on a multilevel multiple membership model and a common method of multilevel analysis (e.g., Goldstein, 2011b, p. 255f) was suggested that takes random noise into account. The decisive factor in this approach is that data is not transformed first and then aggregated but instead is processed top-down. Authors' dispositions (e.g., competency, traits) are assumed to generate publications with citation impact (an assumption made implicitly by the h-index as well). The set of publications of an author represents measurement replications of his or her disposition. If an author repeatedly shows high citation success across various publications, this success cannot be due solely to situational influences or coincidence but must be due also to the person's disposition. In the case of co-authorships, the author can only influence citation success with a respective fraction (e.g., equal weighting). In a multi-membership model, the disposition across all (co-)authorships is measured.

The main empirical results of the study are as follows:

- *The type of weighting*: In fractional counting II, the type of weighting no longer plays a major role. This empirical result confirms a result of an available simulation study (Smith & Beretvas, 2014) that found that different weightings have no influence on the estimation of random effects: "Under the conditions examined here, results indicated that choice of weight pattern did not greatly impact relative parameter bias nor level two residuals' ranks" (Smith & Beretvas, 2014, p. 31).
- *Field normalization*: Despite a relatively homogeneous sample of social scientists with publications in a strongly restricted range of scientific fields, a strong influence of field normalization on citations (50%–60% explained variance) was observed. Thus, field normalization should be an indispensable element of a bibliometric analysis.
- *Intraclass correlation*: Intraclass correlations provide an effective tool to test whether there are any differences between the authors beyond chance. Empirically, high intraclass correlation could be demonstrated, so that the application of multilevel analyses does make sense.
- *Normal distribution*: The question of skewed distributions is certainly a problem at the level of bibliometric raw data (e.g., citations), but at the level of random effects there are approximately normally distributed data, as is the case with other characteristics in psychology or medicine.
- *PPtop10p*: PPtop10p is currently the method of choice in field normalization to achieve robust results. However, dichotomization is accompanied by decreasing ICC and separation of authors with respect to their achievement.

In summary, the following advantages and disadvantages of fractional counting II compared to the common numerical methods can be mentioned:

- *Fractions*: The special technique of fractional counting (e.g., equal weighting, first author weighing) does not play any role (Smith & Beretvas, 2014).
- *Model test:* It is possible to check whether an assumed model fits the data well (e.g., χ^2/df).
- *Duplicates*: Duplicates of publications in the data can be erased.
- *Top-down strategy*: A top-down strategy is favored that defines the object of the study in advance, e.g., the dispositions of authors, which are assumed to be the main cause of the observed citation impact of publications.
- *Bayes estimates*: The estimation of individual parameters for each author considers the reliability of measurements (e.g., number of publications of an author).

- *Discrimination power*: Fractional counting II based on raw citation counts, shows greater power to differentiate between authors than fractional counting II based on percentiles (PPtop10p).
- *Extensions*: The multilevel models presented here can be extended by further levels (e.g., institution of the authors, countries). It is also possible to link network analysis to generate the weights in the multiple-membership model (e.g., Tranmer et al., 2014). Instead of people, journals can also be used to determine the stable contributions of journals. Last but not least, it is also possible to capture interdisciplinarity, in which it is not the authors who form the categories in the multiple-membership model, but the scientific fields. The aim may be to measure the influence of the fields on the citations.

These advantages are offset by disadvantages that may restrict the use of fractional counting II:

- *Big data*: Bibliometric analysis is often characterized by the use of huge data sets. For example, the total data set of Web of Science from 1980 to 2012 (38,508,986 publications) was used to say something about the growth rates of modern science (Bornmann & Mutz, 2015). Such data sets are difficult to process statistically at the level of single publications. However, there are several ways to deal with these huge data sets: First, random samples can be drawn that represent a much more efficient way of collecting data than complete surveys (Mutz, 2016; Williams & Bornmann, 2016). Second, there are now statistical methods that can estimate multilevel models for large data sets (e.g., PROC HPMIXED, SAS). Third, the analyses do not necessarily have to start with single publications but can also refer to sets of publications—for example, the publication set of an author, since summed count data are in turn count data.
- *Transparency:* Statistical methods involve a certain lack of transparency in the estimation of parameters. In principle, however, statistical analyses, if they are well documented, may be replicated by other researchers.
- *Outlier:* Bibliometric data are characterized by the fact that outliers, or extreme values, may occur. Under the assumption of Poisson-distributed count data, extreme values occur for purely statistical reasons. However, their influence is minimized by the logarithmically scaled random effects. Sensitivity analyses are able to check to what extent individual values influence the parameter estimation. Furthermore, a set of publications of an author (measurement replications) is assumed to obtain reliable estimations, so that single highly cited papers add less to the overall disposition of a researcher.

Despite the disadvantages, a more statistical orientation of bibliometrics is necessary in the future, which includes questions of random noise, measurement errors, and model tests (Adler, Ewing, & Taylor, 2008).

Author contributions

Rüdiger Mutz: Conceived and designed the analysis, Collected the data, Contributed data or analysis tools, Performed the analysis, Wrote the paper and Other contribution.

Hans-Dieter Daniel: Conceived and designed the analysis and Wrote the paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.joi.2019.03.007.

References

Adler, R., Ewing, J., & Taylor, P. (2008). *Citation statistics - report* Retrieved from. https://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf

Ajiferuke, I., & Famoye, F. (2015). Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models. *Journal of Informetrics*, *9*(3), 499–513. http://dx.doi.org/10.1016/j.joi.2015.05.001

Altman, D. G. (2006). The cost of dichotomising continuous variables. *BMJ*, *332*(7549), 1080. http://dx.doi.org/10.1136/bmj.332.7549.1080

Assimakis, N., & Adam, M. (2010). A new author's productivity index: p-index. *Scientometrics*, *85*(2), 415–427. http://dx.doi.org/10.1007/s11192-010-0255-z

Austin, P. C., Stryhn, H., Leckie, G., & Merlo, J. (2018). Measures of clustering and heterogeneity in multilevel Poisson regression analysis of rates/count data. *Statistics in Medicine*, *37*(4), 572–589. http://dx.doi.org/10.1002/sim.7532

Bornmann, L., & Daniel, H.-D. (2007). Multiple publication on a single research study: Does it pay? The influence of number of research articles on total citation counts in biomedicine. *Journal of the American Society for Information Science and Technology*, *55*(8), 1100–1107. http://dx.doi.org/10.1002/asi.20531

Bornmann, L., & Daniel, H.-D. (2016). Count regression models in informetrics. *Journal of Informetrics*, *10*(1), 29–30. http://dx.doi.org/10.1016/j.joi.2015.10.003

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics*, *9*(2), 408–418. http://dx.doi.org/10.1016/j.joi.2015.01.006

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics*, *5*(1), 228–230. http://dx.doi.org/10.1016/j.joi.2010.10.009

Bornmann, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics*, *7*(1), 158–165. http://dx.doi.org/10.1016/j.joi.2012.10.001

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222. http://dx.doi.org/10.1002/asi.23329

Bornmann, L., Mutz, R., & Daniel, H.-D. (2013). Multilevel-statistical reformulation of citation-based university rankings: The Leiden ranking 2011/2012. *Journal of the American Society for Information Science and Technology*, *64*(4), 1649–1658. http://dx.doi.org/10.1002/asi.22857

Bornmann, L., Stefaner, M., de Moya Anegon, F., & Mutz, R. (2014). Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualization of results from multi-level models. *Online Information Review*, *38*(1), 43–58. http://dx.doi.org/10.1108/OIR-12-2012-0214

Bornmann, L., Stefaner, M., de Moya Anegon, F., & Mutz, R. (2015). Ranking and mappping of universities and research-focused institutions worldwide: The third release of excellencemapping.net. *COLLNET Journal of Scientometrics and Information Management*, *9*(1), 65–72. http://dx.doi.org/10.1080/09737766.2015.1027090

Bouyssou, D., & Marchant, T. (2016). Ranking authors using fractional counting of citations: An axiomatic approach. *Journal of Informetrics*, *10*(1), 183–199. http://dx.doi.org/10.1016/j.joi.2015.12.006

Boyer, S., Ikeda, T., Lefort, M.-C., Malumbres-Olarte, J., & Schmidt, J. M. (2017). Percentage-based author contribution index: A universal measure of author contribution to scientific articles. *Research Integrity and Peer Review*, *2*(18) http://dx.doi.org/10.1186/s41073-017-0042-y

Burrell, Q. (2007). Hirsch's h-index: A stochastic model. *Journal of Informetrics*, *1*(1), 16–25. http://dx.doi.org/10.1016/j.joi.2006.07.001

Cafri, G., Hedeker, D., & Aarons, G. A. (2015). An introduction and integration of cross-classified, multiple membership, and dynamic group random-effects models. *Psychological Methods*, *20*(4), 407–421. http://dx.doi.org/10.1037/met0000043

Chung, H., & Beretvas, S. N. (2012). The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 185–200. http://dx.doi.org/10.1111/j.2044-8317.2011.02023.x

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics, quantitative methods in library. Documentation and information science*. Amsterdam: Elsevier.

Egghe, L., Rousseau, R., & Van Hooydonk, G. (2000). Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American Society for Information Science*, *51*(2), 145–157. http://dx.doi.org/10.1002/(SICI)1097-4571(2000)51:2<145::AID-ASI6>3.0.CO;2-9

Erlen, J. A., Siminoff, L. A., Sereika, S. M., & Sutton, L. B. (1997). Multiple authorship: Issues and recommendations. *Journal of Professional Nursing*, *13*(4), 262–270. http://dx.doi.org/10.1016/S8755-7223(97)80097-X

Fielding, A., & Goldstein, H. (2006). *Cross-classified and multiple-membership structures in multilevels models: An introduction and review*. Retrieved from University of Birmingham. https://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/cross-classified-review.pdf

Gauffriau, M. (2017). A categorization of arguments for counting methods for publication and citation indicators. *Journal of Informetrics*, *11*(3), 672–684. http://dx.doi.org/10.1016/j.joi.2017.05.009

Goldstein, H. (2011a). Estimating research performance by using research grant award gradings. *Journal of the Royal Statistical Society A (Statistics in Society)*, *174*(1), 83–93. http://dx.doi.org/10.1111/j.1467-985X.2010.00657.x

Goldstein, H. (2011b). *Multilevel statistical models*. Chichester: Wiley.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society A (Statistics in Society)*, *158*(1), 175–177. http://dx.doi.org/10.2307/2983411

Gotthard, A., & Calzolari, G. (2017). Estimating multiple-membership logit models with mixed effects: Indirect inference versus data cloning. *Journal of Statistical Computation and Simulation*, *87*(12), 2334–2348. http://dx.doi.org/10.1080/00949655.2017.1331440

Hagen, N. T. (2010). Harmonic publication and citation counting: Sharing authorship credit equitably—Not equally, geometrically or arithmetically. *Scientometrics*, *84*(3), 785–793. http://dx.doi.org/10.1007/s11192-009-0129-4

Hagen, N. T. (2013). Harmonic coauthor credit: A parsimonious quantification of the byline hierarchy. *Journal of Informetrics*, *7*(4), 784–791. http://dx.doi.org/10.1016/j.joi.2013.06.005

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, *22*(9), 1433–1446. http://dx.doi.org/10.1002/sim.1522

Hilbe, J. H. (2014). *Modeling count data*. Cambridge: University Press.

Hox, J. J., Moerbeck, M., & van de School, R. (2018). *Multilevel analysis: Techniques and applications*. New York: Taylor & Francis.

Joe, H., & Zhu, R. (2005). Generalized Poisson distribution: The property of mixture of Poisson and comparison with Negative Binomial distribution. *Biometrical Journal*, *47*(2), 219–229. http://dx.doi.org/10.1002/bimj.200410102

Kosmulski, M. (2012). The order in the lists of authors in multi-author papers revisited. *Journal of Informetrics*, *6*(4), 639–644. http://dx.doi.org/10.1016/j.joi.2012.06.006

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the Association for Information Science and Technology*, *62*(7), 1370–1381. http://dx.doi.org/10.1002/asi.21534

Lindsey, D. (1992). Production and citation measures in the sociology of science: The problem of multiple authorship. *Social Studies of Science*, *10*(2), 145–162. http://dx.doi.org/10.1177/030631278001000202

Mutz, R. (2016). Some further aspects of sampling: Comment on Williams and Bornmann. *Journal of Informetrics*, *10*(4), 1241–1242. http://dx.doi.org/10.1016/j.joi.2016.09.007

Mutz, R., & Daniel, H.-D. (2015). What is behind the curtain of the Leiden Ranking? *Journal of the Association for Information Science and Technology*, *66*(9), 1950–1953. http://dx.doi.org/10.1002/asi.23360

Mutz, R., Bornmann, L., & Daniel, H.-D. (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PloS One*, *7*(10), e48509 http://dx.doi.org/10.1371/journal.pone.0048509

Mutz, R., Wolbring, T., & Daniel, H.-D. (2017). The effect of the "very important paper" (VIP) designation in Angewandte Chemie International Edition on citation impact: A propensity score matching analysis. *Journal of the Association for Information Science and Technology*, *68*(9), 2139–2153. http://dx.doi.org/10.1002/asi.23701

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2015). Multiplicative versus fractional counting methods for co-authored publications. The case of the 500 universities in the Leiden Ranking. *Jounal of Informetrics*, *9*(4), 974–989. http://dx.doi.org/10.1016/j.joi.2015.10.002

Perianes-Rodriguez, A., Waltman, L., & van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, *10*(4), 1178–1195. http://dx.doi.org/10.1016/j.joi.2016.10.006

Raabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using STATA - volume II: Categorical responses, counts and survival* (3rd ed.). Lakeway Drive, College Station, Texas: Stata Press.

Radicchi, F., & Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E*, *83*(4) http://dx.doi.org/10.1103/PhysRevE.83.046116

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. London: Sage.

SAS Institute Inc. (2011). *SAS/STAT® 9.3 user's guide*. Cary, NC: SAS Institute Inc.

Smith, L. J. W., & Beretvas, S. N. (2014). The impact of using incorrect weights with the multiple membership random effects model. *Methodology*, *10*(1), 31–42. http://dx.doi.org/10.1027/1614-2241/a000066

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics*, *11*(1), 128–151. http://dx.doi.org/10.1016/j.joi.2016.12.002

Tranmer, M., Pallotti, F., & Lomi, A. (2016). The embeddedness of organizational performance: Multiple membership multiple classification models for the analysis of multilevel networks. *Social Networks*, *44*, 269–280. http://dx.doi.org/10.1016/j.socnet.2015.06.005

Tranmer, M., Steel, D., & Browne, W. J. (2014). Multiple-membership multiple-classification models for social network and group dependences. *Journal of the Royal Statistical Society A (Statistics in Society)*, *177*(2), 439–455. http://dx.doi.org/10.1111/rssa.12021

Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H., & Krauss, J. (2007). Author sequence and credit for contributions in multiauthored publications. *PLoS Biology*, *5*(1) http://dx.doi.org/10.1371/journal.pbio.0050018

van Hooydonk, G. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science*, *48*(10), 944–945. http://dx.doi.org/10.1002/(SICI)1097-4571(199710)48:10<944::AID-ASI8>3.0.CO;2-1

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. http://dx.doi.org/10.1016/j.joi.2016.02.007

Waltman, L., & van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, *9*(4), 872–894. http://dx.doi.org/10.1016/j.joi.2015.08.001
Waltman, L., & van Eck, N. J. (2013a). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics*, *96*(3), 699–716. http://dx.doi.org/10.1007/s11192-012-0913-4
Waltman, L., & van Eck, N. J. (2013b). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*(4), 833–849. http://dx.doi.org/10.1016/j.joi.2013.08.002
Williams, R., & Bornmann, L. (2016). Sampling issues in bibliometric analysis. *Journal of Informetrics*, *10*(4), 1225–1232. http://dx.doi.org/10.1016/j.joi.2015.11.004